CERTIFICATE OF MAILING BY "EXPRESS MAIL"

"Express Mail" mailing label number EK171107864US

Date of Deposit: November 1, 2001

I hereby certify that this paper or fee is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date inscribed above and is addressed to the Assistant Commissioner of Patents, Box PATENT APPLICATION, Washington, D.C. 20231.

JOAN PENNINGTON
(Typed or printed name of person mailing paper or fee)

Joan Pennington
(Signature of person mailing paper or fee)

QoS SCHEDULER AND METHOD FOR IMPLEMENTING PEAK SERVICE DISTANCE USING NEXT PEAK SERVICE TIME VIOLATED INDICATION

Field of the Invention

The present invention relates generally to the storage and data networking fields, and more particularly, relates to a QoS scheduler and method for implementing peak service distance using a next peak service time violated indication.

Related Applications

Related United States patent applications by William John Goetzinger, Glen Howard Handlogten, James Francis Mikos, and David Alan Norgaard and assigned to the present assignee are being filed on the same day as the present patent application including:

United States patent application Serial Number _______________, entitled "QoS SCHEDULER AND METHOD FOR IMPLEMENTING QUALITY OF SERVICE WITH AGING TIME STAMPS";

United States patent application Serial Number _______________, entitled "QoS SCHEDULER AND METHOD FOR IMPLEMENTING QUALITY OF SERVICE WITH CACHED STATUS ARRAY";

United States patent application Serial Number _______________, entitled "QoS SCHEDULER AND METHOD FOR IMPLEMENTING QUALITY OF SERVICE ANTICIPATING THE END OF A CHAIN OF

FLOWS";

United States patent application Serial Number ______________, entitled "WEIGHTED FAIR QUEUE HAVING EXTENDED EFFECTIVE RANGE";

United States patent application Serial Number ______________, entitled "WEIGHTED FAIR QUEUE SERVING PLURAL OUTPUT PORTS";

United States patent application Serial Number ______________, entitled " WEIGHTED FAIR QUEUE HAVING ADJUSTABLE SCALING FACTOR"; and

United States patent application Serial Number ______________, entitled "EMPTY INDICATORS FOR WEIGHTED FAIR QUEUES".

Description of the Related Art

Storage and data networks are designed to support the integration of high quality voice, video, and high speed data traffic. Storage and data networking promises to provide transparent data sharing services at high speeds. It is easy to see that rapid movement and sharing of diagrams, pictures, movies, audio, and the like requires tremendous bandwidth. Network management is concerned with the efficient management of every bit of available bandwidth.

A need exists for a high speed scheduler for networking that ensures the available bandwidth will not be wasted and that the available bandwidth will be efficiently and fairly allocated. The scheduler should permit many network traffic flows to be individually scheduled per their respective negotiated Quality-of-Service (QoS) levels. This would give system administrators the ability to efficiently tailor their gateways, switches, storage area networks (SANs), and the like. Various QoS can be set up using combinations of precise guaranteed bandwidth, required by video for example, and limited or unlimited best effort bandwidth for still pictures, diagrams, and the like. Selecting a small amount of guaranteed bandwidth with the addition of some bandwidth from the pool of best effort bandwidth should guarantee that even during the highest peak periods, critical data will be delivered to its application at that guaranteed rate.

A scheduler advantageously may be added to a network processor to enhance the quality of service (QoS) provided by the network processor subsystem.

One of the functions of a QoS scheduler is to limit the best effort bandwidth allocated to a flow based on a peak service distance (PSD) specification. The peak service distance (PSD) specification is a negotiated Quality-of-Service (QoS) level for an individual traffic flow. The QoS scheduler should individually schedule each of multiple flows per their respective assigned PSD specification, even when additional bandwidth is available.

A problem of conventional arrangements results where a flow violates its peak service distance (PSD) specification at the same time that the flow goes empty. Then in conventional arrangements, when another frame for this flow arrives, the flow is scheduled such that the flow may immediately be selected as a winner, that is the flow may immediately identified for servicing. In that case the new frame would be dispatched and the flow would again be empty. This cycle could repeat indefinitely, and the timing could be such that the flow would receive much more bandwidth than specified by its PSD specification.

A need exists for a scheduler and scheduling method for implementing scheduling so that an individual flow does not receive more service than deserved.

Summary of the Invention

A principal object of the present invention is to provide a scheduler and method for implementing peak service distance using a next peak service time violated indication. Other important objects of the present invention are to provide such scheduler and method for implementing peak service distance using next peak service time violated indication substantially without negative effect and that overcome many of the disadvantages of prior art arrangements.

In brief, a scheduler and method are provided for implementing peak service distance using a next peak service time violated (NPTV) indication. A flow scheduled on a best effort or weighted fair queue (WFQ) is identified for servicing and a frame is dispatching from the identified flow. A next PSD time (NPT) for the flow is checked to see if it has been violated. Responsive to identifying the next PSD time (NPT) being violated for the identified flow, a NPTV indicator is set. Alternatively, responsive to identifying the next PSD time (NPT) not being violated for the identified flow, the NPTV indicator is reset. A next PSD time (NPT) value is calculated for the flow. Checking for more frames to be dispatched from the flow is performed. Responsive to identifying no more frames to be dispatched from the flow, the NPTV indicator is utilized to identify a calendar or ring for attaching the flow upon a new frame arrival for the flow.

In accordance with features of the invention, if the NPTV indicator is not set when the flow goes empty, upon a new frame arrival for the flow, the flow is attached to a weighted fair queue (WFQ) ring using a queue distance calculation. If the NPTV indicator is set when the flow goes empty, upon a new frame arrival for the flow, then it is determined if the next PSD time (NPT) value has been passed. If the next PSD time (NPT) value has been passed, then the flow is attached to the weighted fair queue (WFQ) ring using the queue distance calculation. If the next PSD time (NPT) value has not been passed, then the flow is attached to a peak bandwidth service (PBS) calendar using the next PSD time (NPT) value.

Brief Description of the Drawings

The present invention together with the above and other objects and advantages may best be understood from the following detailed description of the preferred embodiments of the invention illustrated in the drawings, wherein:

FIG. 1A is a block diagram illustrating a network processor system including a scheduler for carrying out scheduling methods for implementing peak service distance using next peak service time violated indication of the preferred embodiment;

FIGS. 1B is diagram providing a graphical illustration of various types of QoS algorithms in accordance with the preferred embodiment;

FIG. 2 is a high-level system diagram illustrating the QoS scheduler for carrying out scheduling methods for implementing peak service distance using next peak service time violated indication of the preferred embodiment;

FIG. 3 is a flow chart illustrating prior art steps for attaching a flow after the flow goes empty;

FIG. 4 is a flow chart illustrating exemplary sequential steps for carrying out scheduling methods for implementing peak service distance using next peak service time violated indication for attaching a flow after the flow goes empty of the preferred embodiment; and

FIG. 5 is a block diagram illustrating a computer program product in accordance with the preferred embodiment.

Detailed Description of the Preferred Embodiments

Having reference now to the drawings, in FIG. 1A, there is shown a network processor system generally designated by the reference character 100 including a scheduler 200 for carrying out scheduling methods for implementing peak service distance (consumption of excess bandwidth up to a specified limit) using a next peak service time violated indication of the preferred embodiment. As shown in FIG. 1A, network processor system 100 includes a network processor 102 that executes software responsible for forwarding network traffic. Network processor 102 includes hardware assist functions for performing operations, such as table searches, policing, and statistics tracking. A dataflow 104 serves as the primary data path for transmitting and receiving data flow traffic, for example, via a network interconnect 106 and/or a switch fabric interface 108. Dataflow 104 provides an interface to a large data store memory 110 for buffering of traffic bursts when an incoming frame rate exceeds an outgoing frame rate. An external flow queue memory 112 is coupled to scheduler 200. As network processor performance continues to increase, unique techniques and design solutions enable the QoS scheduler 200 of the preferred embodiment to perform reliably at these high data rates.

Scheduler 200 of the preferred embodiment permits many network traffic flows, for example, 64 thousand (64K) network traffic flows to be individually scheduled per their respective assigned Quality-of-Service (QoS) level. Each flow is basically a one-way connection between two different points. QoS parameters are held in a flow queue control block (FQCB), such as in the external flow queue memory 112. QoS parameters include sustained service distance (SSD), peak service distance (PSD), queue distance (QD), port identification (ID), and the like. There can be, for example, 64 thousand flows and a FQCB for each flow.

FIG. 1B provides a graphical illustration of various types of QoS algorithms. The scheduler 200 provides for quality of service by maintaining flow queues that may be scheduled using various algorithms, such as a set guaranteed bandwidth, or best effort or weighted fair queue (WFQ) with or without a peak bandwidth service (PBS) limit. The best effort or weighted fair queue is limited via the peak service distance (PSD) QoS parameter. The guaranteed bandwidth is set via the sustained service distance (SSD) QoS parameter. A combination of these algorithms provide efficient utilization of available bandwidth. The scheduler 200 supplements the congestion control algorithms of dataflow 104 by permitting frames to be discarded based on per flow queue thresholds.

Referring now to FIG. 2, there is shown a high-level system diagram illustrating the scheduler 200 for carrying out scheduling methods of the preferred embodiment. Scheduler 200 includes a bus interface 202 coupled to a system bus 204 interconnecting modules in the system 100. Chipset messages are exchanged between modules using system bus 204. Messages include flow enqueue requests which add frames to a given flow and read and write requests. Scheduler 200 includes a message buffer 206, such as a first-in first-out (FIFO) message buffer, that stores messages until they are ready to be executed. Scheduler 200 includes a queue manager 208 coupled to the message buffer 206. Queue manager 208 processes the incoming messages to determine what action is required. Queue manager 208 is coupled to calendars and rings block 220 and a memory manager 224. A winner partition 222 arbitrates between the calendars and rings 220 to choose which flow will be serviced next. The memory manager 224 coordinates data reads from and writes to a first and second external static random access memory (SRAM) 226 and 228 and an internal memory array 230.

For a flow enqueue request received by queue manager 208, the flow's FQCB information is retrieved from one of the external SRAM 226 or 228 or internal array 230 and examined to determine if the new frame should be added to an existing frame string for a given flow, start a new frame string, or be discarded. In addition, the flow queue may be attached to a calendar or ring for servicing in the future. Read and write request messages received by queue manager 208 are used to initialize flows.

Port back-pressure from the dataflow 104 to the scheduler 200 occurs via the port status request message originated from the dataflow and applied to the calendar and rings block 220. When a port threshold is exceeded, all WFQ and PBS traffic associated with that port is held in the scheduler 200 and the selection logic of winner partition 222 does not consider those flows as potential winners. When port back-pressure is removed, the flows associated with that port are again eligible to be winners.

Calendars and rings block 220 includes, for example, three calendars (low latency service (LLS), normal latency service (NLS), peak bandwidth service (PBS)) and weighted fair queues (WFQs). The calendars are time based. The weighted fair queues (WFQs) are weight based. The WFQs are also referred to as best effort queues because WFQs can only schedule excess bandwidth and therefore can have no bandwidth guarantee associated with them.

Flows are attached to one or more of three calendars/rings (LLS, NLS, PBS) and one WFQ ring 220 in a manner consistent with its QoS parameters. For example, if a flow has a guaranteed bandwidth component, it is attached to a time based calendar. If a flow has a WFQ component, it is attached to a WFQ ring. A flow may have both a guaranteed component and best effort or WFQ component. The calendars 220 are used to provide guaranteed bandwidth with both a low latency service (LLS) and a normal latency service (NLS) packet rate. Flows are scheduled for service at a certain time in the future. The WFQ rings are used by the weighted fair queuing algorithm. Entries are chosen based upon position in the WFQ rings without regard to time. The WFQ rings are work conserving or idle only when there are no flows to be serviced. A flow set up using a WFQ ring can optionally have a peak bandwidth limit associated with it.

Scheduler 200 performs high speed scheduling, for example, processing 27 Million frames per second (Mframes/second). Scheduling rates per flow for the LLS, NLS and PBS calendars 220 range, for example, from 10 Giga bits per second (Gbps) to 3.397 Thousand bits per second (Kbps). Rates do not apply to the WFQ ring.

SRAM 226 is an external high speed, for example, quad data rate (QDR) SRAM containing flow queue information or flow queue control block (FQCB) information and frame information or frame control block (FCB) information. SRAM 228 is, for example, an optional external QDR SRAM containing flow queue information or flow queue control block (FQCB) depending on the number of flows. Internal array 230 contains for example, 4k FQCB or 64K aging information. Internal array 230 may be used in place of the external SRAM 228 if less than for example four thousand (4K) flows are required and is also used to hold time stamp aging information. Internal array 230 containing FQCB aging information is used with logic that searches through the flows and invalidates expired time stamps.

Queue manager 208 performs the queuing operation of scheduler 200 generally as follows: A linked list or string of frames is associated with each flow. Frames are always enqueued to the tail of the linked list. Frames are always dequeued from the head of the linked list. Flows are attached to one or more of four calendars/rings (LLS, NLS, PBS, WFQ) 220 using the QoS parameters. Selection of which flow to service is done by examining the calendars/rings 220 in the order of LLS, NLS, PBS, WFQ. Then the frame at the head of the selected flow is selected for service. The flow queues are not grouped in any predetermined way to target port. The port number for each flow is user programmable. All WFQ flows with the same port ID are attached to the same WFQ ring. The QoS parameters also apply to the discard flow. The discard flow address is user selectable and is set up at configuration time.

When a flow enqueue request is sent to the scheduler 200, its frame is tested for possible discard using information from the flow enqueue request message and information stored in the FQCB. If the frame is to be discarded then the FQCB pointer is changed from the FQCB in flow enqueue request message to the discard FQCB. Alternatively, the frame is added to the tail end of the FCB chain associated with the FQCB. In addition, the flow is attached if it is not already attached to the appropriate calendar (LSS, NLS, PBS), or ring (WFQ). As time passes, selection logic of winner partition 222 determines which flow is to be serviced (first LLS, then NLS, then PBS, then WFQ). If a port bandwidth threshold has been exceeded, the WFQ and PBS component associated with that port are not eligible to be selected. When a flow is selected as the winner, the frame at the head of the FCB chain for the flow is dequeued and a port enqueue response message is issued to the dataflow 104. If the flow is eligible for a calendar reattach, the flow is reattached to the appropriate calendar (LLS, NLS, PBS) or ring (WFQ) in a manner consistent with the QoS parameters.

In accordance with features of the preferred embodiment, a scheduling method of monitoring flow service is provided so that a flow does not receive more service than deserved. An indicator is used to determine how a new attach should be performed after a flow has gone empty and a new frame for the flow arrives. This indicator is a next PSD time violated (NPTV). If a flow violates its PSD specification at the time the flow goes empty, its NPTV indicator is set to signal the violation. In addition to the indicator being set, if when the new frame arrives not enough time has passed to put the bandwidth for the flow at or below its peak bandwidth specification, then the flow is scheduled directly on the PSD calendar to ensure it does not receive more service than is deserved.

FIG. 3 illustrates a problem resulting from conventional scheduling steps for attaching a flow after the flow goes empty. FIGS. 4A and 4B illustrate scheduling steps for attaching a flow after the flow goes empty in accordance with the preferred embodiment. As shown in FIGS. 3, 4A and

4B, a flow is configured to have weighted fair queue (WFQ) component specified by queue distance (QD) and a peak service distance (PSD). The flow also may or may not have a normal latency service (NLS) component. However, it is assumed for simplicity that the flow has no NLS component because the NLS component is not required to understand the problem solved by the present invention.

When a first frame is attached to the flow, the flow is scheduled on the weighted fair queue (WFQ) ring. If more frames for this flow arrive before the flow is serviced or selected as a winner, these frames are chained onto the flow frame list. At some point the flow is chosen for servicing, that is the flow is picked as the winner. At that time a frame is dispatched from the flow. If the flow has more frames that can be dispatched, then the flow is rescheduled on the WFQ. If the flow has no more frames, that is the flow is empty; then the flow is not rescheduled on the WFQ ring. Whether or not the flow is rescheduled, a next PSD time (NPT) value for the flow is calculated, using the peak service distance (PSD) value and the size of the frame that was just dispatched. The NPT specifies the earliest time that the flow can be serviced again without violating the PSD specification.

As indicated in a decision block 302 after the flow was rescheduled on the WFQ, at some point the flow will again be a winner, and one of its frames will be dispatched. It is determined whether the next PSD time was violated as indicated in a decision block 304. If at that time the frame is dispached, the current time (CT) is greater than or equal to the NPT, then flow has not exceeded its peak bandwidth specification (PSD) and the next PSD time (NPT) was not violated at decision block 304. In this case, a new next PSD time (NPT) is calculated as indicated in a block 306 and it is determined whether the flow has more frames to send as indicated in a decision block 308. When the flow has more frames, the flow is rescheduled on the WFQ ring using a queue distance calculation as indicated in a block 310. Then the sequential steps return to block 302.

If, however, CT is less than NPT, then flow has exceeded its peak bandwidth specification and the next PSD time (NPT) was violated at decision block 304. In this case, there are two ways that the flow might be handled. A new next PSD time (NPT) is calculated as indicated in a block

312 and it is determined whether the flow has more frames to send as indicated in a decision block 314. If the flow has still more frames to be dispatched, then the flow is attached on the PSD calendar using the new next PSD time (NPT) as indicated in a block 316. The flow is scheduled at a time equal to the NPT that was calculated at block 312 for the frame that was just dispatched. This ensures that the flow will not again be serviced until it is at or below its peak bandwidth specification, to avoid violating the PSD specification for the flow. Then the sequential steps return to block 302. If the flow goes empty (there are no more frames to be dispatched), this is identified at decision block 314, and the flow is not rescheduled on either the PSD calendar or on the WFQ ring. This situation, where the flow is not rescheduled on either the PSD calendar or on the WPQ calendar, presents a problem in the prior art in that the flow that violated NPT can receive much more bandwidth than specified by its PSD.

When no more frames for the flow that did not violate NPT are identified at decision block 308, or no more frames for the flow that did violate NPT are identified at decision block 314, checking for a new frame to arrive for this flow is performed as indicated in a decision block 316. When a new frame is identified for either the flow that did not violate NPT or the flow that did violate NPT, then the next PSD time is invalidated as indicated in a block 318 and the flow is attached to WFQ ring using a queue distance (QD) calculation at block 310.

After the new frame for this flow that violated its PSD specification arrives and the flow is scheduled on the WFQ, it is possible that the flow may immediately be selected as a winner. In that case the new frame would be dispatched and the flow would again be empty. This cycle could repeat indefinitely, and the timing could be such that the flow would receive much more bandwidth than specified by its PSD specification.

Referring now to FIGS. 4A and 4B, there are shown exemplary sequential steps for carrying out scheduling methods for implementing peak service distance for attaching a flow after the flow goes empty using a next peak service time violated indication of the preferred embodiment. Scheduler 200 as shown in FIGS. 4A and 4B solves the prior art scheduling problem of FIG. 3 that is possible when a flow violates its peak service distance (PSD) specification at the same time that the flow goes empty. An indicator called next PSD time violated (NPTV) is provided by the preferred embodiment to rectify the prior art scheduling problem described above.

Referring now to FIG. 4A, a flow is picked as a winner as indicated in a decision block 402 and one of its frames is dispatched. It is determined whether the next PSD time (NPT) was violated as indicated in a decision block 404. As described above, the next PSD time (NPT) was not violated at that time the frame is dispached, if the current time (CT) is greater than or equal to the NPT, then flow has not exceeded its peak bandwidth specification (PSD). If the next PSD time (NPT) was violated at that time the frame is dispached, then the next PSD time violated (NPTV) indicator is set for the flow that has exceeded its PSD specification as indicated in a block 406. A new next PSD time (NPT) is calculated as indicated in a block 408. Then the sequential steps continue following entry point B in FIG. 4B.

When the next PSD time (NPT) was not violated at that time the frame is dispached, then the next PSD time violated (NPTV) indicator is reset for the flow that has not exceeded its PSD specification as indicated in a block 410. A new next PSD time (NPT) is calculated as indicated in a block 412. Then the sequential steps continue following entry point C in FIG. 4B.

Referring now to FIG. 4B, following entry point B checking for more frames to send for the flow that has exceeded its PSD specification is performed as indicated in a decision block 414. When the flow has more frames to be dispatched, then the flow is attached on the PSD calendar using the new next PSD time (NPT) as indicated in a block 416. This ensures that the flow will not again be serviced until it is at or below its peak bandwidth specification, to avoid violating the PSD specification for the flow. Then the sequential steps return following entry point A in FIG. 4A.

In FIG. 4B following entry point C, checking for more frames to send for the flow that has not exceeded its PSD specification is performed as indicated in a decision block 418. When the flow has more frames to be dispatched, then the flow is attached on the WFQ ring using the queue distance calculation as indicated in a block 420. Then the sequential steps return following entry point A in FIG. 4A.

In accordance with features of the preferred embodiment, when the flow goes empty, then the NPTV indicator is used to determine how a new attach should be performed when a new frame for the flow arrives. When no more frames for the flow that did violate NPT are identified at decision block 414, or no more frames for the flow that did not violate NPT are identified at decision block 418, checking for a new frame to arrive for the flow is performed as indicated in a decision block 422. When a new frame is identified for either flow, then it is determined whether the NPTV indicator is set as indicated in a decision block 424.

If the NPTV indicator is off, then the flow did not violate its PSD specification when it went empty. The flow, therefore, is scheduled on the WFQ ring using the queue distance calculation at block 420. Then the sequential steps return following entry point A in FIG. 4A.

If the NPTV indicator is set or on, checking whether NPT has aged out or is invalid is performed as indicated in a decision block 426. When the NPTV indicator is set or on, and NPT is not valid because the flow has been aged out by the FQCB aging array 230 or the current time (CT) is greater than or equal to NPT, then the flow is scheduled on the WFQ ring using the queue distance calculation at block 420. The flow that violated its PSD specification when it went empty is attached to the WFQ ring because enough time has already passed such that the bandwidth for the flow is again at or below the peak bandwidth or PSD specification. Then the sequential steps return following entry point A to block 402 in FIG. 4A.

Otherwise, if the NPTV indicator is on at block 424, and NPT is valid where the CT is less than NPT, then the flow is scheduled on the PSD calendar at NPT at block 416. This is because the flow violated its PSD specification when it went empty, and not enough time has passed to put the bandwidth for the flow at or below its peak bandwidth or PSD specification.

Referring now to FIG. 5, an article of manufacture or a computer program product 500 of the invention is illustrated. The computer program product 500 includes a recording medium 502, such as, a floppy disk, a high capacity read only memory in the form of an optically read compact disk or CD-ROM, a tape, a transmission type media such as a digital or analog communications link, or a similar computer program product. Recording medium 502 stores program means 504, 506, 508, 510 on the medium 502 for carrying out scheduling methods for implementing peak service distance using a next peak service time violated indication of the preferred embodiment in the system 100 of FIG. 1A.

A sequence of program instructions or a logical assembly of one or more interrelated modules defined by the recorded program means 504, 506, 508, 510, direct the scheduler 200 for implementing peak service distance using a next peak service time violated indication of the preferred embodiment.

While the present invention has been described with reference to the details of the embodiments of the invention shown in the drawing, these details are not intended to limit the scope of the invention as claimed in the appended claims.